102(a)

B3

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
G01N 33/566, 33/551, 33/552, 33/544

A1

(11) International Publication Number: **WO 98/43088**

(43) International Publication Date: 1 October 1998 (01.10.98)

(21) International Application Number: PCT/US98/05644

(22) International Filing Date: 24 March 1998 (24.03.98)

(30) Priority Data:
60/041,688 24 March 1997 (24.03.97) US

(71) Applicant *(for all designated States except US)*: SMITHKLINE BEECHAM CORPORATION [US/US]; One Franklin Plaza, Philadelphia, PA 19103 (US).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: KAYNE, Paul, S. [US/US]; 111 Cirak Avenue, East Norriton, PA 19403 (US).

(74) Agents: GIMMI, Edward, R. et al.; SmithKline Beecham Corporation, Corporate Intellectual Property, UW2220, 709 Swedeland Road, P.O. Box 1539, King of Prussia, PA 19406–0939 (US).

(81) Designated States: CA, JP, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: METHOD OF PRODUCING A SUBTRACTION LIBRARY

(57) Abstract

A method of producing a subtraction library using a collection of defined sequences is described. The method involves providing a surface containing a collection consisting of known nucleic acid sequences, which is subsequently contacted with a library containing undefined sequences under appropriate hybridization conditions. Any non–hybridized DNA is recovered and sequenced. The resulting sequenced, non–hybridizing DNA forms a subtraction library which contains sequences which were present in the library, and which differ from the sequences of the collection. Also described is a subtraction library prepared according to the method of the invention. Methods are also provided for making and using subtracted probes.

① array of known
② hybridize w/ unknown
③ sequence any that do not hybridize

differs from claim 1 what is immobilized on the array

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

METHOD OF PRODUCING A SUBTRACTION LIBRARY

This application claims priority to United States Provisional Application Serial Number 60/041,688, filed March 24, 1997.

Field of the Invention

The present invention relates generally to the field of generation of cDNA libraries, and more specifically, to methods of generating subtraction libraries.

Background of the Invention

Methods have been described for obtaining information about gene expression and identity using so called "high density DNA arrays" or grids. *See, e.g.*, M. Chee *et al.*, Science, 274:610-614 (1996) and other references cited therein. Such gridding assays have been employed to identify certain novel gene sequences, referred to as Expressed Sequence Tags (EST) (*Adams et a.*, Science, 252:1651-1656 (1991)). A variety of techniques have also been described for identifying particular gene sequences on the basis of their gene products. For example, *see* International Patent Application No. WO91/07087, published May 30, 1991. In addition, methods have been described for the amplification of desired sequences. For example, *see* International Patent Application No. WO91/17271, published November 14, 1991.

Currently available subtraction techniques remove unwanted sequences from a given library. In one approach, a large number of unknown genes are used to drive the subtraction library to remove the unknown genes from a library of interest. *See, e.g.*, J. Love and P. Deininger, BioTechniques, 11(1):88-92 (1991). However, while this technique is useful in removing a large number of genes from the library, little is known about the genes in the resulting subtraction library, other than their source. In another common approach to subtraction, a small number of known genes (*e.g.*, less than 100) are used to drive the subtraction to remove these sequences from the library of interest. This approach allows one to control which genes are removed, but current methods only permit this approach to be used with a small number of genes and the resulting subtraction library contains a significant amount of genes which are not desired.

Accordingly, there exists a need for more efficient methods for producing subtraction libraries. Also needed are more efficient methods for screening for novel pharmaceutical reagents.

Summary of the Invention

In one aspect, the present invention provides a method of producing a subtraction library utilizing a collection containing known or defined sequences. The method involves the steps of providing a surface having a collection of defined nucleic acid sequences and allowing this surface to come into association with a library containing undefined nucleic acid sequences under conditions which permit hybridization. The non-hybridized nucleic acid sequences are recovered, isolated, and form the subtraction library. The subtraction library is characterized by containing the undefined sequences from the second library which are not present in the first library of defined sequences.

In another aspect, the present invention provides a subtraction library produced according to the method of the invention.

In yet another aspect, the present invention provides a method of rapidly screening a library containing undefined sequences for the presence of known or defined sequences using the method steps described herein.

In still another aspect, the present invention provides a method of rapidly screening a library containing undefined sequences using a polynucleotide probe from a known or defined sequences using the method steps described herein.

Other aspects and advantages of the present invention are described further in the following detailed description of the preferred embodiments thereof.

Detailed Description of the Invention

The present invention provides a method of producing a subtraction library utilizing a collection of known or defined sequences. The method involves providing a surface having immobilized thereon a collection consisting of defined or known nucleic acid sequences, which is subsequently contacted with a library containing unknown or undefined sequences under appropriate hybridization conditions. Any non-hybridized polynucleotide, preferably DNA, is recovered as the subtraction library which contains sequences which were present in the library, and which differ from the sequences of the collection. Advantageously, the method of the invention permits a large number of defined

sequences to be used to drive a subtraction. Thus, this method permits the efficient production of a subtraction library whose content can be readily controlled.

Also described is a subtraction library prepared according to the method of the invention.

I. Definitions

Several words and phrases used throughout this specification are defined as follows:

As used herein, the term "gene" refers to the genomic nucleotide sequence from which a cDNA sequence is derived. The term gene classically refers to the genomic sequence, which upon processing, can produce different RNAs.

By "gene product," it is meant any polypeptide sequence, peptide or protein, encoded by a gene.

As used herein, "collection comprising known sequences," refers to any ordered set of nucleotide sequences, including RNA and DNA sequences, which may be in the form of plasmids, cDNA, PCR products, genes, gene fragments, DNA fragments, oligonucleotides and the like. Preferably, such known sequences within the collection have been previously sequenced and/or are of known origin. Desirably, this collection contains a large number of sequences, e.g., as many as 100,000 - 200,000 members where the sequences are genes, or as many as 1,000,000 if the members of the collection include oligonucleotides. However, the number of sequences in the invention may be varied as desired and are not a limitation on the present invention. This collection may contain sequences drawn from a number of different sources, including a variety of libraries. For example, the collection may be drawn from one or more tissue source libraries of a member of the mammalian species, *e.g.*, a human. Desirably, the human is healthy; however, libraries derived from a diseased or impaired individual may also be utilized. Further, other mammals of interest include, without limitation, a non-human primate, a rodent, and a canine. In a particularly preferred collection, defined sequences are present in a single copy.

When utilizing gene sequences in the methods of the invention, it is not necessary that the full-length sequence of the gene be known. Rather, all that the methods of the invention require is that the portion of the gene sequence that renders it unique is known, which is approximately 17 base pairs. By "known nucleic acid sequence," it is meant that the sequence is reasonably unique and contains no redundancies or repeats.

The term "library" includes, but is not limited to, plasmid libraries, RNA libraries, DNA libraries such as those containing PCR products from genomic libraries, cDNA libraries, oligonucleotide libraries and known sequences. Methods for the construction of such libraries are well known by those skilled in the art. A library may be adjusted to minimize the number of complete genes present in a single insert to approximately one gene. Techniques for this adjustment are well known to the skilled artisan.

"Isolated" means altered "by the hand of man" from its natural state; i.e., that, if it occurs in nature, it has been changed or removed from its original environment, or both. For example, a polynucleotide or a polypeptide naturally present in a living animal in its natural state is not "isolated," but the same polynucleotide or polypeptide separated from the coexisting materials of its natural state is "isolated," as the term is employed herein. For example, with respect to polynucleotides, the term isolated means that it is separated from the chromosome and cell in which it naturally occurs.

As used herein, the term "solid support" refers to any substrate which is useful for the immobilization of a plurality of defined materials (i.e., sequences) derived from a library by any available method to enable detectable hybridization of the immobilized polynucleotide sequences with other polynucleotides in the sample. Among a number of available solid supports, one desirable example is the support described in International Patent Application No. WO91/07087, published May 30, 1991. Examples of other useful supports include, but are not limited to, nitrocellulose, nylon, glass, silica and Pall BIODYNE C membrane. It is also anticipated that improvements yet to be made to conventional solid supports may also be employed in this invention.

The term "grid" means any generally two-dimensional structure on a solid support to which the defined materials of a library are attached or immobilized.

As used herein, the term "predefined region" refers to a localized area on a surface of a solid support on which is immobilized one or multiple copies of a particular amplified gene region or sequence and which enables hybridization of that clone at the position, if hybridization of that clone to a sample polynucleotide occurs.

By "immobilized," it is meant to refer to the attachment of the genes or other nucleic acids to the solid support. Means of immobilization are known and conventional to those of skill in the art, and may depend on the type of support being used.

By "label" as used herein is meant any conventional molecule which can be readily attached to or incorporated onto RNA or DNA and which can produce a detectable signal, the intensity of which indicates the relative amount of hybridization of the RNA to the

DNA fragment or oligonucleotide on the grid. Preferred labels are fluorescent molecules or radioactive molecules. A variety of well-known labels can be used.

The term, "subtraction library," as used herein refers to a library that is highly enriched for novel nucleotide sequences, including novel genes and gene fragments, among other sequences.

II. **The Collection of Known Sequences**

In the practice of this method, one or more grids is prepared, so that each grid carries on its solid surface nucleotide sequences from a collection of defined sequences immobilized on the surface. Desirably, this collection is as defined above and these nucleotide sequences are, e.g., genes, gene fragments, other DNA fragments, or oligonucleotide sequences.

Nucleotide sequences from the selected collection are gridded onto a surface of a solid support. Desirably, the nucleotides are in the form of, but are not limited to, DNA, which are put down on the surface in an amount between about 100 pg to about 1000 ng per spot depending on substrate. In a further preferred embodiment the amount of polynucleotide used is between about 10 ng to about 100 ng per spot. However, RNA may be utilized in similar amounts. Although not required, it may be desirable to provide duplicate or multiple coverage of the genes or other nucleotide sequences on the surface. The nucleotide sequences include, but are not limited to, individual clones spotted onto and grown on a surface of the solid support; or plasmid clones isolated from said library, PCR products derived from the plasmid clones, or oligonucleotides derived from sequencing of the plasmid clones, which are immobilized to the surface of the solid support.

Numerous conventional methods are employed for immobilizing these nucleotide sequences to surfaces of a variety of solid supports. *See, e.g.*, Affinity Techniques, Enzyme Purification: Part P, Methods in Enzymology, Vol. 34, ed. W.B. Jakoby, M. Wilcheck, Acad. Press, NY (1971); Immobilized Biochemicals and Affinity Chromatography, Advances in Experimental Medicine and Biology, Vol. 42, ed. R. Dunlap, Plenum Press, NY (1974); U.S. Patent 4,762,881; U.S. Patent No. 4,542,102; European Patent Publication No. 391,608 (October 10, 1990); or U.S. Patent No. 4,992,127 (November 21, 1989).

Although not required, it may be desirable to immobilize the gene or other nucleic acid sequences in an array, such that the sequences are placed at predefined locations or regions on the surface. Knowing how the sequences are arrayed gives the methods of the invention a level of efficiency beyond that which is capable using the prior art methods. This is an important feature of the invention and it allows ready access to the desired

Array

sequence, and preferably its related clone and associated data, following screening. One desirable method for attaching these materials to a solid support is described in International Application No. PCT/US90/06607 (published May 30, 1991). Briefly, this method involves forming predefined regions on a surface of a solid support, where the predefined regions are capable of immobilizing the materials. The method makes use of binding substrates attached to the surface which enable selective activation of the predefined regions. Upon activation, these binding substances become capable of binding and immobilizing the materials derived from the collection of defined sequences.

Any solid substrates suitable for binding nucleotide sequences on the surface thereof for hybridization and methods for attaching nucleotide sequences thereto may be employed by one of skill in the art according to the invention. Currently, however, the preferred surface is glass. As with other solid substrates, methods for depositing and binding nucleotide sequences to a glass surface are well known to those of skill in the art. *See, e.g.*, L. A. Chrisey, *et al.*, "Covalent Attachment of Synthetic DNA to Self-Assembled Monolayer Films", Nucleic Acids Res., 24:3031-3039 (1996); Silicon Compounds: Register and Review (United Chemical Technologies, Inc., Bristol, Pennsylvania, 1993). In an alternative embodiment of the invention, the surface may be beads. Because it is not necessary that the surface be flat, a vertical surface is contemplated to be within the scope of the invention. Such a surface may have steps, ridges, kinks, terraces, and the like.

III. The Library Comprising Undefined Sequences

Once the grid surface containing the immobilized sequences from the collection is prepared, it is allowed to associate with sequences derived from a library containing undefined sequences under suitable hybridization conditions. Such a library may include sequences of unknown origin and/or unknown nucleotide sequences.

The library which provides the source of the unknown or undefined genes, gene fragments, or other nucleic acid sequences may be a random cDNA library obtained using known techniques. Alternatively, a library of genes from a selected organ or tissue, or a mixed set of RNAs, may be the source of the sequences. Suitably, for use in the method of the invention, RNA is isolated and reverse transcribed to cDNA using standard procedures for molecular biology such as those disclosed by *Sambrook et al.*, MOLECULAR CLONING, A LABORATORY MANUAL, 2nd Ed; Cold Spring Harbor Laboratory Press, Cold Spring Harbor Lab Press, Cold Spring Harbor, NY 1989. The cDNA library is then

constructed in accordance with procedures described by Fleischmann *et al.*, *Science*, 269:496-512 (1995). For the purposes of the present invention, the library comprising the undefined sequences may be a library, as is defined above. Most desirably, however, the library is a plasmid library.

In a preferred embodiment, the source of the unknown or undefined sequences is a phagemid library composed of human DNA, and preferably, human cDNA. This embodiment is particularly advantageous for isolation of the non-hybridizing sequences which compose the subtraction library produced by the method of the invention. Such a phagemid library can be prepared using conventional techniques. *See, e.g.*, Sambrook *et al.*, *supra*. Suitable vectors for use in this system include pBluescript [Stratagene, La Jolla, CA] and pUC118 [*J. Vieira and J. Messing*, Methods in Enzymology, 153:3 (1987)]. Other suitable vectors are well known and may be readily selected by one of skill in the art.

As discussed above, however, other suitable techniques and vectors may be utilized to prepare the undefined library. For example, a phage library may be produced using a vector such as M13; however, other suitable vectors are known in the art.

Optionally, these undefined sequences may be labeled to permit detection of DNA which hybridizes to the immobilized sequences. Known conventional methods for labeling the sequences may be used. For example, fluorescence, radioactivity, photoactivation, biotinylation, energy transfer, solid state circuitry, and the like may be used in this invention.

Desirably, the single stranded DNA is isolated from the library containing the undefined sequences using conventional techniques. For example, where the library is a phagemid library, single stranded copies of the library are packaged into particles using a suitable helper phage. The resulting phage particles are then isolated by convention techniques, which typically involve centrifugation and precipitation. Such techniques for isolation of single stranded DNA are well known to those of skill in the art and are not a limitation on the present invention. Although less desirable, double stranded DNA may be utilized in the method of the invention.

IV. Hybridization to the grids

The isolated polynucleotide, preferably DNA (*e.g.*, cDNA), obtained from the undefined library is permitted to come into association with the immobilized sequences derived from the defined driver library under conditions which permit hybridization.

Preferably, hybridization takes place under stringent conditions, *e.g.*, conditions such that only sequences which are more than about 90% identical will remain hybridized throughout the procedures. However, if desired, other less stringent conditions may be selected. For example, less stringent conditions may be desired in order to increase hybridization, thereby decreasing the size of the subtraction library. Thus, a first hybridization may be performed at very low stringency, permitting a significant amount of hybridization and producing a very small subtraction library. Subsequent washes increased or increasing stringencies may then be used to control the size and content of the subtraction library.

Techniques and conditions for hybridization at selected stringencies, such as those described herein, are well known in the art. *See, e.g.*, Sambrook *et al.*, Molecular Cloning. A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1989).

It is preferred that multiple rounds of hybridization are carried out, preferably using the conditions set forth elsewhere herein. It most preferred that the number of rounds of hybridization be between about 3 and 6.

V. Isolation of Non-Hybridized DNA

Following completion of hybridization, the "undefined" sequences which did not hybridize to the immobilized sequences on the grid surface are recovered and purified away from the hybridization solution using standard techniques. *See, e.g.*, Sambrook *et al.*,*supra*.

In a preferred embodiment, wherein the DNA is produced using the phagemid system, the isolated, non-hybridized DNA is converted to double-stranded plasmid by synthesizing the second strand of the non-hybridized DNA. The resulting plasmid is propagated under suitable conditions in an appropriate cell. Suitable host cells may be readily determined by one of skill in the art, taking into consideration the type of plasmid utilized. Desirably, however, the host cells are selected from among bacterial cells. Currently, the preferred bacterial host is an *E. coli* strain. Following cell culture, the DNA from any resulting colonies is isolated using conventional techniques.

In another embodiment, by utilizing labelled sequences, the method of the invention also permits rapid identification of sequences which the library has in common with the immobilized sequences from the collection. Particularly, according to the invention, upon hybridization one can readily detect labelled sequences, and thereby identify those sequences which are being removed from the library and will not be

contained in the subtraction library. The detected sequences are those which are common to the undefined library and the collection of defined sequences.

The collection of DNA isolated from the colonies forms the subtraction library of the invention. This subtraction library is characterized by containing DNA sequences present in the library containing undefined sequences, but excludes those sequences which were present in the collection of defined sequences which was immobilized on the grid surface. Techniques for maintaining these subtraction libraries are well known to those of skill in the art.

The DNA in the resulting subtraction library may be sequenced using standard protocols [Sambrook *et al.*, *supra* or ABI Prizm sequencing kit, Foster City, CA] or utilized for a variety of other purposes.

VI. Subtraction library

Thus, the present invention provides a subtraction library produced according to the method of the invention. This subtraction library provides a source of novel nucleotide sequences, including novel genes and gene fragments, among other sequences. These sequences and, particularly the genes and gene fragments, may be useful in screening drug candidates. The information generated thereby can be used in the pharmaceutical industry to identify new drugs.

Further, these sequences may be employed in conventional methods to produce isolated proteins or peptides encoded thereby. To produce a protein or peptide of this invention, the polynucleotide sequences, preferably DNA, of a desired gene of the invention or portions thereof identified by use of the methods of this invention are inserted into a suitable expression system. In a preferred embodiment, a recombinant molecule or vector is constructed in which the polynucleotide sequence encoding the protein or peptide is operably linked to a heterologous expression control sequence permitting expression of the human protein. Numerous types of appropriate expression vectors and host cell systems are known in the art for mammalian (including human), insect, yeast, fungal and bacterial expression.

The transfection of these vectors into appropriate host cells, whether mammalian, bacterial, fungal or insect, or into appropriate viruses, results in expression of the selected proteins. Suitable host cells, cell lines for transfection and viruses, as well as methods for construction and transfection of such host cells and viruses are well-known. Suitable

methods for transfection, culture, amplification, screening and product production and purification are also known in the art.

In one embodiment, the nucleotides and proteins or peptides encoded thereby which have been identified by this invention can be employed as diagnostic compositions useful in the diagnosis of a disease or infection by conventional diagnostic assays. For example, a diagnostic reagent can be developed which detectably targets a nucleotide sequence or protein of this invention in a biological sample of an animal. Such a reagent may be a complementary nucleotide sequence, an antibody (monoclonal, recombinant or polyclonal), or a chemically derived agonist or antagonist. Alternatively, the nucleotides of this invention and proteins or peptides encoded thereby, fragments of the same, or complementary sequences thereto, may themselves be used as diagnostic reagents. These reagents may optionally be detectably labeled, for example, with a radioisotope or colorimetric enzyme. Selection of an appropriate diagnostic assay format and detection system is within the skill of the art and may readily be chosen without requiring additional explanation by resort to the wealth of art in the diagnostic area.

Additionally, genes and proteins or other sequences identified according to this invention may be used therapeutically. For example, nucleotides or proteins or, peptides identified using the subtraction library of the invention may serve as targets for the screening and development of natural or synthetic chemical compounds which have utility as therapeutic drugs. Alternatively, compounds which inhibit expression of a gene or protein are also believed to be useful therapeutically. In addition, compounds which enhance the expression of genes essential to an organism may also be used.

Conventional assays and techniques may be used for screening and development of such drugs. For example, a method for identifying compounds which specifically bind to or inhibit proteins encoded by these nucleotide sequences can include simply the steps of contacting a selected protein or gene product with a test compound to permit binding of the test compound to the protein; and determining the amount of test compound, if any, which is bound to the protein. Such a method may involve the incubation of the test compound and the protein immobilized on a solid support. Still other conventional methods of drug screening can involve employing a suitable computer program to determine compounds having similar or complementary structure to that of the gene product or portions thereof and screening those compounds for competitive binding to the protein. Identical compounds may be incorporated into an appropriate therapeutic formulation, alone or in combination with other active ingredients. Methods of formulating therapeutic

compositions, as well as suitable pharmaceutical carriers, and the like are well known to those of skill in the art.

Accordingly, through use of such methods, the present invention is believed to provide compounds capable of interacting with these genes (or other nucleotide sequences), or encoded proteins or fragments thereof, and either enhancing or decreasing the biological activity, as desired. Thus, these compounds are also encompassed by this invention.

VII. Subtracted Probes

Further, the present invention provides probes produced by subtraction according to the method of the invention. These probes may be obtained, for example, using the following method, as well as using methods described elsewhere herein. Subtracted probes may be made by subtracting known genes from a mixture of unknown polynucleotides, preferably DNA. Following rounds of subtraction as set forth in the invention, the remaining polynucleotides are the probes for probing unknown polynucleotides, such as by the hybridization methods of the invention. These probes may be labeled using well known methods, such as colorimetric labeling or radiolabeling. These subtraction probes provide a source of novel nucleotide sequences, including novel genes and gene fragments, among other sequences, that are particularly useful to detect unknown polynucleotide sequences, especially in mixtures and libraries, gridded or in solution. These probe sequences and, particularly the genes and gene fragments, may be useful in screening target candidates for drug screening. The information generated thereby can be used in the pharmaceutical industry to identify new drugs.

Subtracted probes, may also be used to prime polynucleotide amplification reactions, such as PCR. To achieve this, the subtracted probes could be made using degenerate polynucleotides sequences comprising a priming site on at least one end of the molecule. Two priming sites may also be added to the molecule, one at each end; and these may be the same or different sequences. The priming site, for example a known sequence between about 5 and 40 nucleotides in length, may be added to the subtracted probes using ligase. These subtracted probes may be used as primers for amplification in solution or on a solid support. In a preferred method, these primers are hybridized to a collection of polynucleotides on a solid support. Following hybridization, double stranded polynucleotide, such as double stranded DNA, could be produced by amplification, preferably by PCR. The polynucleotide primers useful for amplification comprise

sequences that is complementary to the priming site, or sites, where there is more than one and they are different. The amplified fragments may be cloned, for example, using any known method or as described herein, such as using the zero blunt end kit from InVitrogen (Carlsbad, CA).

Further, these probe sequences may be employed in conventional methods to obtain genes by hybridization which genes may be used produce isolated proteins or peptides encoded thereby. Methods described elsewhere herein may be used to produce a protein or peptide of this invention.

In one embodiment, the probe nucleotides identified by this invention can be employed as diagnostic compositions useful in the diagnosis of a disease or infection by conventional diagnostic assays as described elsewhere herein or known in the art.

These examples illustrate the preferred methods of the invention. These examples are illustrative only and do not limit the scope of the invention.

Example 1 - Preparation of a Subtraction Library

A. Immobilization of Collection of Known cDNA Sequences

The driver collection, which contains defined cDNAs, is gridded onto a solid glass surface as follows. The collection is engineered into the pBluescript Vector (Stratagene, La Jolla, CA) and inserts are recovered via PCR using vector-specific primers. The inserts containing the cDNA are deposited on the glass surface via microcapillaries and attached using standard techniques [Silicon Compounds: Register and Review (United Chemical Technologies, Inc., Bristol, Pennsylvania, (1993)].

B. Generation of Undefined cDNA Library

The undefined cDNA library from the desired source is constructed using the SuperScript Plasmid system (Life Technologies, Gaithersburg, MD) according to manufacturer's protocol with the exception that a modified vector is substituted. Briefly, pUC118 [*J. Vieira and J. Messing*, Methods in Enzymology, 153:3 (1987)] is modified to contain the desired cloning sites, and to remove sequences present in the pBluescript multiple cloning site to avoid spurious hybridization. Deletion of the undesired sequences may be performed using Quick Change Mutagenesis kit (Stratagene, La Jolla, CA), according to manufacturer's protocol or another conventional method. The vector has

previously been engineered to contain sequences which permit isolation of single stranded DNA packaged as a bacteriophage, *i.e.*, a phagemid vector.

C. **Subtraction**

Single stranded DNA is isolated from the library to be subtracted (i.e., the cDNA library of B) using a helper phage according to methods described in Sambrook *et al.*, cited above. The single stranded library is hybridized to the gridded library of Part A. using 4x SSC, 42°C, 16-48 hours, at 20 µls. After the appropriate hybridization period, the hybridization solution is recovered, and precipitated using standard techniques (*Sambrook et al.,supra*). In a preferred embodiment, multiple rounds of hybridization are carried out, preferably using the conditions of this Example 1C. It is most preferred that the number of rounds of hybridization be about 4.

Alternatively, hybridization may be performed at a lower temperature, e.g., 37°C. After hybridization, the hybridization solution is first collected and then treated as below. The grid is then washed with 2X SSC at 37°C, the wash collected, and then precipitated. The DNA is recovered as described below. These steps may then be repeated at a higher temperature, e.g., 65°C, and then with 0.2 X SSC at both temperatures.

Following hybridization, the precipitated DNA is converted to double stranded DNA using standard procedures (*Sambrook et al.*, cited above.). Briefly, a vector-specific oligonucleotide is hybridized to the library, followed by synthesis of the second strand by *E. coli* DNA polymerase in the presence of T4 DNA ligase to complete the reaction.

The resulting double stranded DNA is electroporated into an appropriate *E. coli* host strain (e.g., DH5alpha, Life Sciences Technology, Gaithersburg, MD) and the resulting colonies are harvested, forming the subtraction library.

All publications and references, including but not limited to patents and patent applications, cited in this specification are herein incorporated by reference in their entirety as if each individual publication or reference were specifically and individually indicated to be incorporated by reference herein as being fully set forth. Any patent application to which this application claims priority is also incorporated by reference herein in its entirety in the manner described above for publications and references.

The above description fully discloses the invention, including preferred embodiments thereof. Modifications and improvements of the embodiments specifically disclosed herein are within the scope of the following claims. Without further elaboration, it is believed that one skilled in the art can, using the preceding description, utilize the present invention to its fullest extent. Therefore, the examples provided herein are to be construed as merely illustrative and are not a limitation of the scope of the present invention in any way. The embodiments of the invention in which an exclusive property or privilege is claimed are defined as follows.

What is claimed is:

1. A method of producing a nucleic acid subtraction library comprising the steps of:

(a) providing a surface having immobilized thereon a collection comprising known nucleic acid sequences;

(b) hybridizing a library of nucleic acid sequences comprising undefined sequences with the surface; and

(c) isolating the non-hybridized nucleic acids.

2. The method according to claim 1, wherein the surface is glass.

3. The method according to claim 1, wherein the nucleic acids isolated from the library are DNA.

4. The method according to claim 3, wherein the DNA is single-stranded.

5. The method according to claim 1, wherein the nucleic acids of the collection are DNA.

6. The method according to claim 5, wherein the DNA is cDNA.

7. The method according to claim 2, wherein the library is produced using a phagemid vector.

8. The method according to claim 7, wherein the library is produced using the pUC118 vector.

9. The method according to claim 7, wherein the non-hybridized DNA is isolated by the steps comprising:

(a) synthesizing the second strand of the non-hybridized DNA to produce active phagemid DNA;

(b) propagating the phagemid; and

(c) isolating at least one clone.

10. The method according to claim 9, wherein the phagemid is propagated in bacterial cells.

11. The method according to claim 2, wherein the library is produced using a phage vector.

12. The method according to claim 11, wherein the non-hybridized DNA is isolated by the steps comprising: synthesizing the second strand of the non-hybridized DNA to produce active phage DNA, propagating the phage, and isolating plaques.

13. The method according to claim 1, wherein the library is produced from a mixed set of RNAs.

14. The method according to claim 1, wherein the library is a human cDNA library.

15. A subtraction library produced according to the method of claim 1.

16. A sequence isolated from a subtraction library produced according to the method of claim 1.

17. An isolated protein produced by expression of a sequence of claim 15.

18. A method of rapidly screening a library containing undefined sequences for the presence of defined sequences comprising the steps of:

(a) providing a surface having a collection comprising known nucleic acid sequences;

(b) hybridizing nucleic acid sequences isolated from a library containing undefined sequences with the surface, wherein said sequences are associated with a label; and

(c) detecting hybridized nucleic acid sequences.

19. The method according to claim 18, wherein the nucleic acid sequences isolated from the library are DNA.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/05644

A. CLASSIFICATION OF SUBJECT MATTER
IPC(6) : G01N 33/566, 33/551, 33/552, 33/544
US CL :436/518, 524, 527, 528
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 436/518, 524, 527, 528

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	AASHEIM et al. A simple subtraction method for the isolation of cell-specific genes using magnetic monodispersed polymer particles. BioTechniques. April 1994, Vol. 16, No. 4, pages 716-718, 720 and 721, see entire document, especially abstract.	1, 3, 13, 15, 16 ---------- 4-8, 14, 17
X --- Y	US 5,589,339 A (HAMPSON ET AL) 31 December 1996 (31/12/96), see entire document, especially column 5, line 21-column 6, line 31 and column 10, lines 40-68.	1, 3-7, 13, 15-16 ---------- 2, 8-12, 11, 14, 17

TP248.13. B55

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
09 JUNE 1998

Date of mailing of the international search report
27 JUL 1998

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231
Facsimile No. (703) 305-3230

Authorized officer
JOSEPH W. RICIGLIANO
Telephone No. (703) 308-1235

Form PCT/ISA/210 (second sheet)(July 1992)★